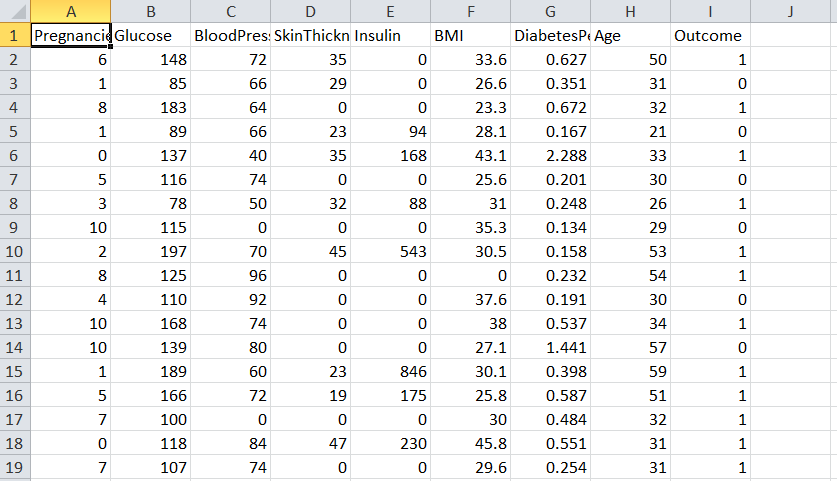
SD03Q06 – Kaalishwar.R – 1832027

Diabetes:

Problem Statement:

The diabetes dataset is a collection of records containing the information about a group of people having diabetes or not based on some features or attributes. We are going to predict the outcome of the patients and predict the output based on a machine learning algorithm. The algorithms used are KNN algorithm and Random Forest classifier.

Dataset:



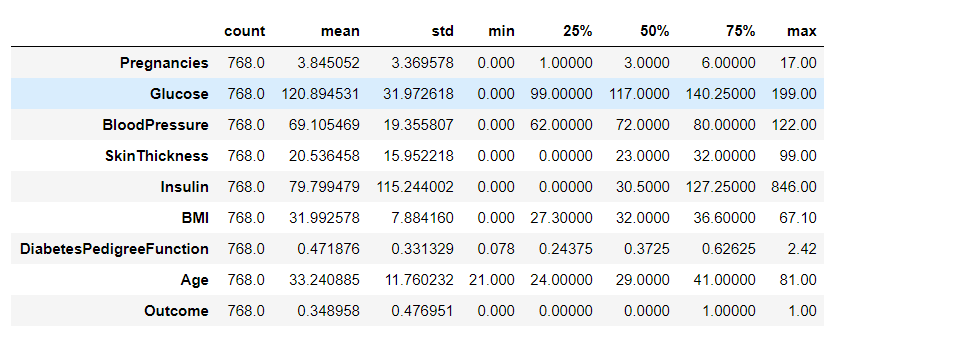
Data preprocessing:

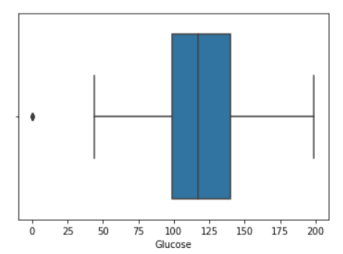
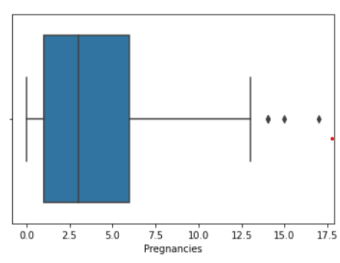
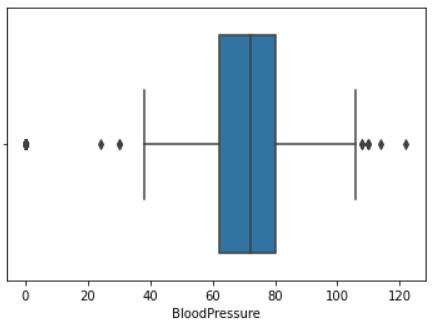
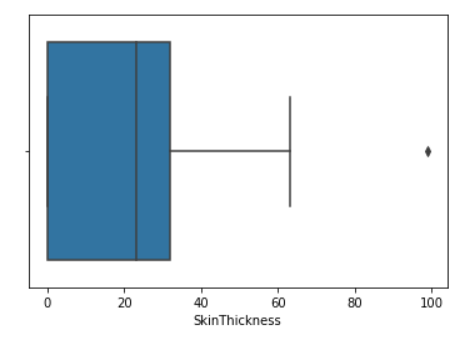
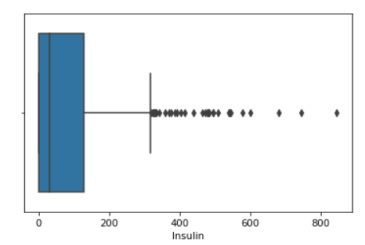
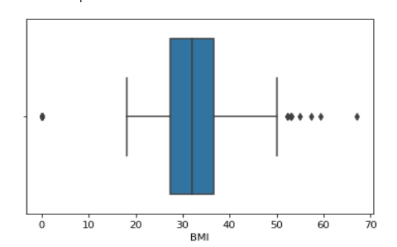
The data contains some zero or missing values. So it can affect the prediction results which can lead to misinterpretation of the results. The zeroes have been replaced with replaced values based on the nature of the distribution of the data , it can’t affect the prediction of the model.

KNN Algorithm:

* Select the number K of the neighbours
* Calculate the Euclidean distance of K number of neighbours. The Euclidean distance shows that the distance between the data points and the distribution of the data points.
* Take the K nearest neighbours as per the calculated Euclidean distance.
* Among these k neighbours, count the number of the data points in each category.
* Assign the new data points to that category for which the number of the neighbour is maximum.

The data has been pre-processed and the missing values have been replaced. The exploratory data analysis has been made so that it describes the dataset.

Now the outliers may affect the prediction so the plots have been made to identify the outliers and the outliers have been removed.

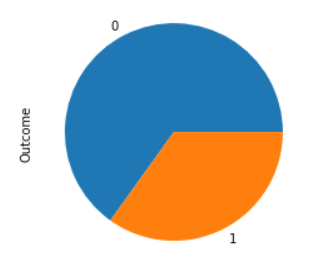
    
  

The outliers have been removed.

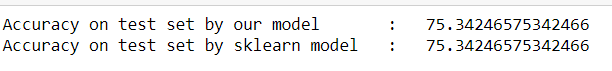
i)KNN Classification:

The KNN model is one of the simplest model easy to understand based on the supervised learning technique. The KNN algorithm can be used for both regression and classification but it is more suitable for classification. It is a non-parametric algorithm, since it does not make any assumption on underlying data.

The model build based on the KNN algorithm gives the prediction of the diabetes. The output attribute has been checked if it is biased or not using the pie chart. The output has been biased to the output 0 which indicates that the person doesn’t have the diabetes. The outliers have been removed.



The data has been separated into training and test set so that the prior one will train the model and the latter will test the data. The neighbours are calculated and K has been calculated.



The accuracy of the KNN model is 75.3424%.

ii)Random Forest Classifier:

Random forest is a supervised learning algorithm which is used for both classification and regression. Random forests is a collection of decision trees on data samples and gets the prediction or result from each tree and select the best prediction.

Algorithm:

* First, start with the selection of random samples from a given dataset.
* Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
* In this step, voting will be performed for every predicted result.
* At last, select the most voted prediction result as the final prediction result.

The model has been built on the training set and it has been tested by the test set.

The model gives the accuracy of 82.23350%.



Conclusion:

The random forest classifier gives the highest accuracy of 82.23% compared with other models. The KNN model has the accuracy of 75.34%.